METHODS AND ALGORITHMS OF ANALYZING SYLLABUSES FOR EDUCATIONAL PROGRAMS FORMING INTELLECTUAL SYSTEM

[1]D. KAIBASSOVA, [2]L. LA, [3]A. SMAGULOVA, [4]L. LISITSYNA, [5]A. SHIKOV, [6]M. NURTAY

[1] Doctoral student, L.N.Gumilyov Eurasian National University, specialty Information Systems, Nur-Sultan, Kazakhstan

[2] Docent, L.N.Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan

[3]Docent, Karaganda State Technical University, Karagandy, Kazakhstan

[4] Professor, ITMO University, Saint Petersburg, Russia

[5]Associate Professor, ITMO University, Saint Petersburg, Russia

[6] Engineer, Karaganda State Technical University, Karagandy, Kazakhstan

E-mail: [1]dindgin@mail.ru, [2]lira_la@mail.ru, [3]asemgul_sss@mail.ru, [4]lisizina@mail.ifmo.ru, [5]shik-off@mail.ru

ABSTRACT

This article reviews using methods of intellectual data analysis for educational program formation in the context of determining the sequence of studying disciplines in the direction by consideration. The model of the forming educational programs that satisfy given competencies is described on the basis of text documents processing through their vector representations. Proposed model performs clustering of text documents taking into weights coefficient of individual words in the corpus. The article succinctly describes the developed software application that allows extract information from text documents, process, analyze, and visualize data. Testing was carried out according to data obtained from 350 syllabuses of disciplines for conformity with 120 competencies in the areas of IT-specialists training. This research solves the issues of intellectual support for the educational programs disciplines of higher education with a view to diminish the complexity of developing new educational programs and improve the quality of academic content.

Keywords: *Educational Program, Information Extraction, Vectorization, Text Mining, Cosine Similarity, Hierarchical Clustering.*

1. INTRODUCTION

The implementation of information technologies for formation of educational programs at universities is a requirement not only for students but also for employers as it is possible to identify qualification specifications for candidates for various vacancies. For the formation of educational programs It becomes important to analyze the content of academic discipline in concomitance with requirements of educational and professional standards taking into account the demand of the labor market for professional disciplines and the direction of educational program. In this regard, the design of educational programs provides procedure for creating a basis of ***professional*** competencies based on professional standards.

Due to the introduction of educational process automation at universities the functionality of information systems which allows to develop curricula, disciplines work programs, and methodical documentation. But there are no such kind of software applications on the market that allow regardless of the subject area carry out a comparative analysis of the content, goals and training results of educational programs and courses with a view in order to further actualization and formation of educational programs taking into account the latest requirements of educational and professional standards. The authors [1] have proposed a general structure of the system which could provide informational maintenance for lecturers in solving this problem where one of the main goals is not only to disentangle the process of enhancement, but also focus on the coveted result of higher education – professional skills and qualities of student. It was reached by implementing the requirements designated in professional standards, and the actual job instructions given by real employers and by analyzing the relationship between these requirements and the learning outcomes declared in academic documents themselves [1].

The development of educational programs to identify the totality of disciplines corresponding to the presented competencies in one of the main tasks of educational institutions. There are many works to clarify such tasks where machine learning methods are used. In this work [2] applied such eminent approaches of data analysis as Hierarchical Cluster Analysis, Principal Component Analysis (PCA), Non-negative Matrix Factorization to average the weights of competencies. The authors [3] in their work suggest using the description, academic results and structure of the training course using the “mean average precision” metric to calculate and compare text similarities. They put it in an application for the various documents ranking according to a query document and evaluating the results using an average accuracy indicator. In another work [4] the authors consider miscellaneous approaches to the semantic analysis of academic courses through their vector representations, including the TF-IDF algorithm.

Potentials of integrating data mining components into a wide range of application software is relevant. One such component is natural language processing. At the same time some predicaments arise in the computer processing of a natural language because natural languages determined not by rules, but by feature extractions. The process of extracting features or vectorization (converting a document into a vector representation) is the first step towards analyzing a natural language. A prominent task in the field of intellectual analyzing educational data is to extract the structure of knowledge, for instance, in works from ontology [5-8] which describes the framework and relations between didactic units of training courses and programs. Particularly, the language analysis components based upon a contemporary text data examination infrastructure [9-13, 28] which includes collections of techniques and methods that combine tools for working with strings, lexical resources, computational linguistics, and machine learning algorithms convert natural language data into machine form and contrary (e.g. machine translation).

Existing approaches to the intellectual support of the educational programs formation based on ontological models of systems inspired by knowledge and rules, heuristic algorithms for automated curriculum development, expert assessment methods and cognitive maps do not allow for effective accounting and operational tracking of changes both in the labor market and educational content space. In turn, the formation of ontological models, systems of rules and precedents by experts for all available subject areas of educational programs preparation is an extremely laborious process requiring the involvement of a representative team of experts in each of the subject areas to ensure the necessary accuracy. This study is dedicated to solving the important and urgent task of intellectual support for the educational programs formation, which has a high complexity when processing large volumes of poorly structured information in a short period of time under constant modifications. The analysis of current publications and practical developments revealed that the proposed approach was not described earlier and was not used to intellectual support of educational programs formation. In this work classical statistical models and methods of text mining and information retrieval were used, which allow to solve the problem of identifying a set of disciplines that meet the requirements in professional and educational standards.

Nowadays, there are two ways to analyze text in natural language: linguistic analysis which based on extracting the meaning of the text by its semantic structure and statistical analysis based upon extracting the meaning of the text by the frequency distribution of words in the text. In this case, we used the statistical method of text analysis, based on counting the frequency of words occurrence.

Initially it is necessary to build a vector model of documents selected by information search queries represented by vectors. In event of searching for a document by query [9, 26], the query is also represented as a vector of the same space and it is possible to calculate the compliance of documents with the query. This approach is focused on finding documents by their content. This work uses Vector Space Model (VSM) – a mathematical model for representing text data in a single space R^n. Each document is represented as a vector $(a_1, a_2, ..., a_n)$, where a_i is the weight of the i-th word. To define the a_i weight tf (term frequency) was used, i.e. the ratio of number of times a word appears in a document compared to the total number of words in that document.

This article is the result of a study to solve the intellectual support’s problem for the process, which identifies the appropriate working curricula of disciplines for the formation of professional competencies of educational programs in order to improve the quality of educational content. The subject of the study is a working curriculum (syllabus), which is defined as a combination of data characterizing the learning outcomes and the discipline's content. During the experimental work, 350 syllabuses of disciplines were analyzed for compliance with 120 competencies in the areas of training IT specialists.

2. VECTOR MODEL OF SUBMISSION OF DOCUMENTS FOR THE INTELLECTUAL SYSTEM

In meeting linguistic tasks of text processing various approaches and methods for converting textual information into sets of numerical data are possible which will be used to extract knowledge, in particular, compare texts and identify matches in them, tools of automatic linguistic analysis are needed [14–16, 28]. The algorithm for constructing a vector model can be written in the form of the following sequential steps:

1. Text preprocessing. At the pre-processing stage, all images, tables, sentences with formulas, information about authors and bibliographic references are deleted from the source text. Only the thematic content of the discipline is retrieved. The end result of this stage is a body from a collection of documents.

2. Text conversion. This step is a filtering of "stop words", special characters and numbers, from which subsequently a "bag of words" is formed and prepared for building a vector model.

3. Assessment of the document by external features. The importance of this stage is to determine whether a document belongs to the relevant competency using a vector model.

The principal merit of the vector model is the ability to search and rank documents by their affinity in the vector space. To obtain the vector of weights for a document, it is necessary to index the documents. Thus, indexing a document refers to the process of mapping the text of a document to its logical presentation. Indexing can be represented in three stages:

1. Terms extraction. At this stage, methods are used to search and select the most substantial terms in the documents corpus.

2. Terms weighting. This stage determines the significance of the term for the selected document.

3. Dimensionality reduction. It is the process of reducing vector space.

Terms extraction. It is the process of breaking text into simpler objects and it also called feature extraction. The result of this process is a set of terms T, which will be used to obtain the weight characteristics of the document.

Lexical analysis. In turn, lexical analysis is the first step in extracting terms. At this stage, all non-letter characters are eliminated, for example punctuation marks, numbers, brackets, and html tags, etc.

Removing stop words. Words without any independent semantic meaning charge are usually called stop words. Stop words include prepositions, conjunctions, and pronouns [17]. To reduce the dimension of the term space the indexer does not take into account stop words and removes them during analysis. Stop words also markedly affect keyword selection. If they are not removed, they would clog many terms as they are often found in the text.

Lemmatization and stemming. Reduction each word in a document to normal form is called lemmatization [17]. For reduction a text in Russian to normal form apply, first of all, for nouns and adjectives – nominative case, singular, masculine; secondly, for verbs and participles – an infinitive verb.

Stemming is a discarding the changeable parts of words, mainly endings. This technology is plainer and does not require the storage of words dictionary or a large set of rules. The technology is based upon the rules of language morphology [17].

All terms contained in the processed texts (including taxonomic ones) are added to a single set after processing, which does not contain reduplicative words. The indicated set is usually called the "bag of words". Suppose $|T| = n$.

Let's describe a model for the formation of an educational program that satisfies given competencies. The problem statement can be formulated as follows:

Suppose there is a collection of syllabuses $D = (d_1, d_2, ..., d_n)$ and a glossary of terms $S = (s_1, s_2, ..., s_m)$ from a competency base that interacts with a user request. Consider a document from the collection of syllabuses $d_i \in D$ and present it as a vector from space R^n. Then this vector will have the form:

$$d_i = (tf_{i1}, tf_{i2}, ..., tf_{in})$$

where tf_{ij} is the number with which the word from the query occurs in the document. The document in the vector model is considered as an unorderedset of items.

Need to get a matrix

$$X = \{x_{ij}\}$$

where $x_{ij} = tf(s_j, d_i)$ is the frequency of terms found in the document.

Consider the sequence of actions for organizing a search by s_j for in a document d_i.

1. A text document is selected from a collection of documents;
2. The stop words have removed from the text;
3. Given the morphology of words, the frequency of occurrence of each term is calculated;
4. Terms have ranked in descending order of their occurrence frequency;
5. A term-document matrix is formed.

Thus, a term-document matrix is created in which the rows correspond to the documents from the collection and the columns conform to the terms [18]. As a result, the null rows correspond to syllabuses to which no competency conforms, because they are not considered for further analysis, so such words are deleted. They may also be zero columns that define uncovered competencies, in such a situation it is necessary to add syllabuses formed in this area of knowledge.

For the next stages to identify a group of syllabuses that are similar in meaning data mining and clustering methods are used. Consider the most common measures for evaluating the importance of words and text fragments in processed text documents. To build a vector model weights of identical words are calculated using the TF-IDF method. We will describe this procedure in more detail.

Let $T = \{t_i\}$ represents the set of words that are found in a collection of documents D. Each document $d_k \in D$ is associated with a vector $\overrightarrow{d_k} = (x_0, x_1, ..., x_n)$ n-dimensional space ($n = |T|$), where x_i is the weight of word $t_i \in T$ in the document d_k calculated using the *tf-idf* method.

The majority of works consider the frequency of terms using in text across only a single document. For the statistical analysis of the frequency of words using in a document against a collection of documents (corpus) it is convenient to use the TF-IDF method (Term Frequency - Inverse Document Frequency). Applying this method it can be obtained the weight of each term in relation to the body. For this, weight for the number of times that it appears in this document is added to the term, and reduced for the number of the rest of documents in which this term is used.

TF (Term Frequency) is the numerical value of a given word occurrence in the current document. It is calculated by the formula:

$$TF = \frac{n_i}{\sum n_k} \qquad (1)$$

where n_i is the number of a given word's occurrences, n_k is the total number of words in the document.

IDF (Inverse Term Frequency) is a numerical value that shows how often this word appears in all source of documents. Calculation formula:

$$IDF = \log\left(\frac{D}{d_i}\right) \qquad (2)$$

where D is the total number of documents, and d_i are the documents in which the given word occurs [18].

The final value of the TF-IDF coefficient is equal to the product of above factors

$$TFIDF = TF \cdot IDF \qquad (3)$$

Words with a high frequency within a given document and with a low frequency within a whole set of documents gain more weight. To calculate TF, a vector of normalized words is used. Each word becomes a key in the map, and the number of occurrences becomes a value.

That completes the preparatory phase and let's start the work of the clustering algorithm. The goal of document clustering is to automatically identify groups of semantically similar documents amongst a given fixed set of documents [9, 14, 16]. To conduct cluster data analysis, similarity measures are used [19]. Sneath, Sokal divided the measures of similarity into four types: correlation coefficients; distance measures; associativity coefficients and probabilistic similarity coefficients [17, 19]. It should be marked that groups are formed only on the basis of pairwise documents descriptions similarity. This is a necessary condition for calculating the distance between two vectors, which, in turn, makes it possible to compare texts by comparing the vectors representing them in any metric (Euclidean distance, cosine measure, Manhattan distance, Chebyshev distance, etc.), i.e., producing cluster analysis.

At the cluster analysis stage a measure of cosine similarity between vectors was used. Cosine similarity is a measure of similarity between two vectors of the pre-Hilbert space, which is used to measure the cosine for an angle between them. If two feature vectors A and B are given, then the cosine similarity cos (θ) can be represented using the scalar product and the norm [18]:

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \qquad (4)$$

The cosine similarity value of the two documents varies in the range from 0 to 1, since the frequency of the term (TFIDF weights) cannot be negative. The angle between two term frequency vectors cannot be greater than 90°. One of the reasons for the traction of cosine similarity is that it is effective as an estimated measure, especially for sparse vectors, since only nonzero measurements must be taken into account [21].

In concordance to the structure of many clusters, clustering algorithms can generate flat clustering or hierarchical clustering. Hierarchical clustering implies the presence of a nested clusters tree. The advantage of hierarchical over flat algorithms is that hierarchical clustering allows to get more information about the selection of documents and gives the user the opportunity to consider different levels of thematic organization of the collection [22]. Classical approaches to the construction of hierarchical clustering are agglomerative and divisive clustering algorithms. In hierarchical clustering using agglomerative algorithms, objects are gradually combined into ever larger clusters. Thus, from the configuration, when each object is a separate cluster, one cluster is obtained containing all the objects. On the other hand, when using divisive algorithms smaller clusters are obtained from larger clusters. At the same time, clusters from individual objects are obtained from one cluster containing all the objects in the selection.

When constructing the clustering using agglomerative method the object is initially considered as a separate cluster. For singleton clusters the distance function is naturally determined:

$$R(\{x\},\{x'\}) = p(x,x')$$

Then the merging process starts. At each iteration, instead of a pair of the closest clusters U и V a new cluster is created $W = U \cup V$. The distance from a new cluster W to any other cluster S is calculated from the distances $R(U,V), R(U,S), R(V,S)$ that should be known by this moment:

$$R(U \cup V, S) = \alpha_U R(U,S) + \alpha_V R(V,S) + \\ + \beta R(U,V) + \gamma |R(U,S) - R(V,S)| \qquad (5)$$

where $\alpha_U, \alpha_V, \beta, \gamma$ are the numerical parameters. This universal formula for calculating intercluster distances was supposed by Lance and Williams in 1967 [23].

There are a number of ways for calculating the distances $R(W,S)$ between clusters W и S, and for each of them the compliance with the Lance-Williams formula for certain combinations of parameters is proved. Works [24, 25] in which the agglomerative and divisional text clustering algorithms were studied, it was established that the average distance is the best metric for agglomerative methods. In turn, the average distance is calculated by the formula:

$$R(W,S) = \frac{1}{|W||S|} \sum_{w \in W} \sum_{s \in S} p(w,s);$$
$$\alpha_U = \frac{|U|}{|W|}, \alpha_V = \frac{|V|}{|W|}, \beta = \gamma = 0 \qquad (6)$$

In this method, the distance between two different clusters is computed as the average distance between all pairs of objects in them. The **scipy** library was used to implement the above methods **cluster**.

3. IMPLEMENTATION OF METHODS AND ALGORITHMS FOR ANALYSIS OF DOCUMENTS FOR THE INTELLECTUAL SYSTEM FOR THE FORMATION OF EDUCATIONAL PROGRAMS

To conduct experiments in order to obtain a vector model of syllabuses for the formation of educational programs, a software application was developed using the algorithms above. Let's turn to the description of the software application's main blocks.

The main blocks of the application **(**fig. 1**)** are a data preprocessing unit, a visualization unit, and a vector model formation unit. The data preprocessing unit is intended for transforming a sequence of document terms into an n-dimensional vector space, i.e. maps the text of a document to its logical presentation. The result of the first block is the indexed document.

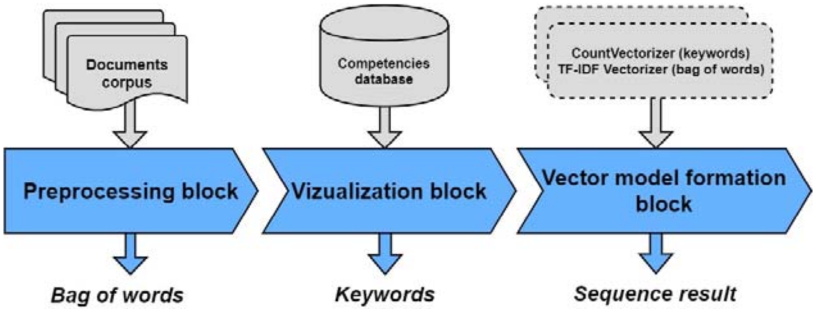


Figure 1: The main blocks of software application

The visualization block represents the selected documents according to external characteristics. External signs are keywords built on the basis of competencies. The significance of the developed software application consists in determining the necessary disciplines which form the professional competencies that should be set by the user at the initial stage of working with the application.

To implement text mining methods, the following structural model was constructed and illustrated in Figure 2. The structural model reveals the basic processes of software application. The text corpus was formed from a collection of working educational programs of disciplines, which in turn is input information. During the data preprocessing the text is normalized using the methods of deleting various characters including numbers and punctuation marks, as well as eliminating many auxiliary words imported as stop words from NLTK library for working with the natural language. Subsequently, a “bag of words” for further analysis is received has been received after normalization. At the next stage a vector model is formed based on the frequency and weight analysis of the use of words in documents. The end result is a vector that sets the sequence for the study of academic disciplines to relevant competencies.

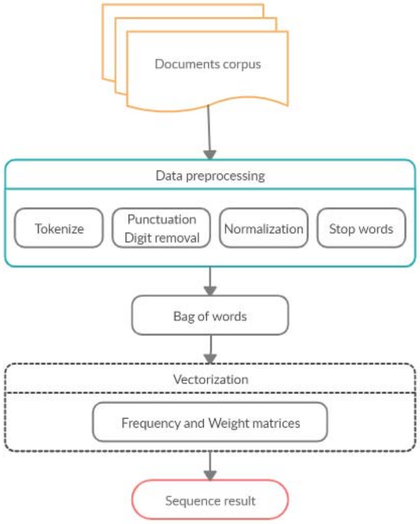


Figure 2: Structural model of the system

The software application was developed applying basic architectural solutions that made it easier to develop the interface, make the system more flexible and resistant to changes, as well as dynamically change data. The structural pattern that describes the way to build the structure of our application the areas of responsibility and the interaction of each of the parts in this structure allows the MVC concept (Model-View-Controller). In the MVC paradigm, user input, modeling of the outside world, and visual feedback from the user are separated from each other, and processed by three types of objects specialized to perform their task [26]. The interaction between objects and classes adapted to solve the design problem in this context with a reasonable degree of refinement was determined by the following class structures and the inheritance hierarchy in Figure 4.

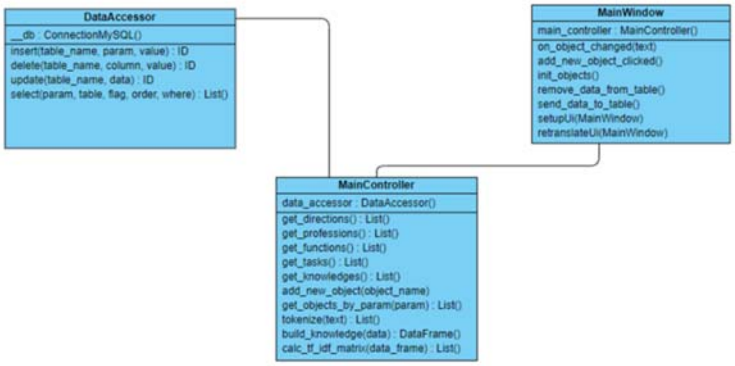

DataAccessor
__db : ConnectionMySQL()
insert(table_name, param, value) : ID
delete(table_name, column, value) : ID
update(table_name, data) : ID
select(param, table, flag, order, where) : List()
MainWindow
main_controller : MainController()
on_object_changed(text)
add_new_object_clicked()
init_objects()
remove_data_from_table()
send_data_to_table()
setupUi(MainWindow)
retranslateUi(MainWindow)
MainController
data_accessor : DataAccessor()
get_directions() : List()
get_professions() : List()
get_functions() : List()
get_tasks() : List()
get_knowledges() : List()
add_new_object(object_name)
get_objects_by_param(param) : List()
tokenize(text) : List()
build_knowledge(data) : DataFrame()
calc_tf_idf_matrix(data_frame) : List()

Figure 3: Classes diagram

The figure demonstrates the class diagram of the software application. The main class is MainWindow which implements the work of the interface part. This class is represented by the main_controller field to interact with the User using such controls as text fields, drop-down lists, and tables. This class serves as the View level and is responsible for obtaining the necessary data from the model and sends it to the user, so it defined through the association relationship with the following MainController class. Moreover, the MainController class has a data_accessor object which acts as an intermediary between the user interface and the database. The MainController class also performs the business logic functions, precisely, it performs operations on a vector model of data representation. It uses methods of data access level (DataAccessor class) and view (MainWindow) to implement the necessary actions, for example, such as adding directions to educational programs, labor functions and competencies. The DataAccessor class is a data access level class that provides a connection to the database. In addition, this class contains methods for extracting, adding, and processing the necessary data.

To form an educational program using the developed software application, it is recommended:

1. Determine the desired area of professional activity after graduation;

2. Set the competencies required for this field of activity.

One of the main elements is a competency model based on a database that represents all the fundamental aspects of the management and formation of educational programs.

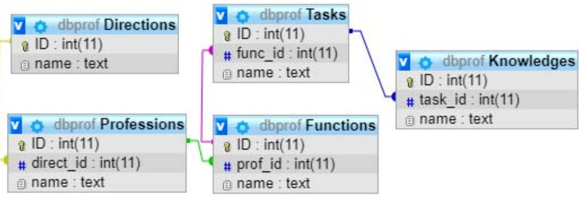


Figure 4: Database scheme

The database structure consists 5 main tables referred to each other by a parent-child relationship. Let us briefly describe the fields of these tables. The Directions table is designed to store information about the educational programs directions: the identifier of the direction (id), the name of the direction (name). The Professions table contains information about the professions in the relevant areas: identifier (id), field (direct_id) as a foreign key, name of the profession (name). It should be noted here that foreign is a column or combination of columns whose values correspond to the primary key in another table, in this case, the values of the direct_id column (foreign key) correspond to the values of the id column (primary key) from the Directions table. The remaining Functions, Tasks, Knowledges tables are created in the same way. Building a database depend on a relational DBMS would enable to use the built-in full-text search, regulate dependency relationships between competencies, and also implement some linguistic algorithms using stored procedures and functions.

The interaction of the database with the visualization unit is organized using a fetch command, which, in turn, performs the function of converting the input text into a matrix, the values of which are the number of occurrences of this key (word) in the text.

4. THE OUTCOMES OF EXPERIMENTAL WORKS

To determine the group of syllabuses characterized by common properties, as well as to find groups of similar syllabuses in the sample, a cluster analysis of documents was used. The cluster method is a multidimensional statistical procedure that collects data containing information about a sample of objects, and then organizes the objects into relatively homogeneous groups. For clustering text documents, you must perform the following operations:

1) A selection of objects (syllabuses) for clustering;

2) Definition of the variables set by which the selected syllabuses are evaluated;

3) A similarity measure calculation between syllabuses;

4) Cluster analysis applying to create groups of similar objects (clusters).

The objects selection for clustering involves the formation of keywords vector from a professional competencies database selected by the user. The selection was carried out as follows (queries listing 1):

```
     SELECT name FROM directions
     SELECT id FROM directions WHERE name='ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ'
     SELECT name FROM professions WHERE direct_id=1
     SELECT id FROM professions WHERE name='РАЗРАБОТЧИК ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ'
     SELECT name FROM functions WHERE prof_id=7
     SELECT  id  FROM  functions  WHERE  name='Подготовка  к  разработке  программного
обеспечения'
     SELECT name FROM tasks WHERE func_id=8
     SELECT id FROM tasks WHERE name='Анализ требований к программному обеспечению'
     SELECT name FROM knowledges WHERE task_id=10
```

Queries listing 1: Keywords Selection

As a result of the keywords vector formation $T = \{t_1, t_2, ..., t_n\}$ competencies were selected. ['Жизненный цикл программного обеспечения', 'Программное обеспечение и его функциональные возможности', 'Методы выявления требований к программному обеспечению']. The proposed algorithm compares words from the text of a documents collection with words from a vector of keywords and builds a frequency matrix. Then it finds among them the words with the maximum number of occurrences in the text. Elements of this matrix (table 1) contain weights that take into account the frequency of each term’s use in each document.

Table 1: Fragment of a frequency matrix

№	Document	программный	принцип	организация	стандарт	проектирование	процедура	Обеспечение	шаблон
1	IT-инфраструктура.txt	8.0	3.0	7.0	3.0	3.0	NaN	NaN	NaN
2	WEB-программирование.txt	2.0	3.0	1.0	NaN	NaN	NaN	NaN	NaN
3	АСДП.txt	1.0	NaN	NaN	NaN	NaN	1.0	NaN	NaN
4	Автоматизированный бухгалтерский учет и финансо...	2.0	NaN	2.0	NaN	NaN	NaN	1.0	NaN
…	. . .								

348	Администрирование информационных систем.txt	1.0	NaN	NaN	NaN	NaN	1.0	NaN	NaN
349	Интеллектуальный анализ данных.txt	1.0	NaN	NaN	1.0	NaN	NaN	NaN	NaN
350	Компьютерные сети.txt	1.0	5.0	1.0	NaN	1.0	NaN	NaN	NaN

According to the criterion for the specific gravity of keywords, the most appropriate disciplines were selected. The specific gravity serves as an indicator of the frequency phenomenon that is the number of elements in the total volume of the population. It should be noticed that zero lines correspond to syllabuses that do not correspond to any competence, since they are not considered for further analysis therefore, such lines are deleted. There may also be zero columns that define competencies not covered, in which case it is necessary to add syllabuses formed in this area of knowledge. We get a frequency matrix of 10 rows, in which the rows correspond to syllabuses from the collection, the columns correspond to the terms that were selected in concordance with the proposed algorithm (competency). All in all, thematic related documents can be seen in the frequency graph (fig. 5).

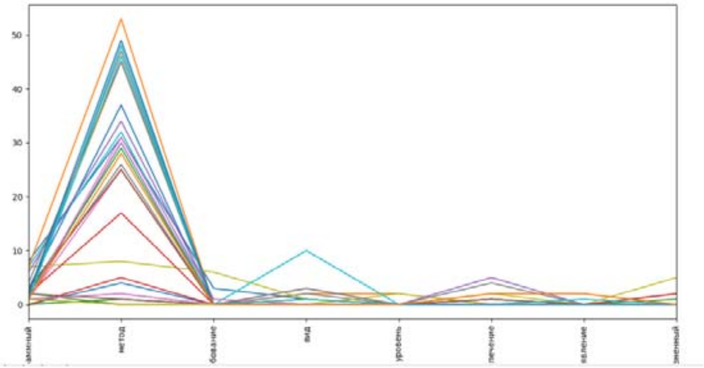


Figure 5:Frequency graph

Summing up the implementation phase, intermediate tasks were solved relating to the methods of Text Mining. 350 documents that were contained in the collection, after the work were distributed as follows:

Table 2: Fragment of the weights matrix

	алгоритм	анализ	безопасность	ветвление	взаимодействие	вид
IT-инфраструктура 2019.txt	0	0,06680	0,01028	0	0,03111	0,00835
АСДП.txt	0,15297	0	0	0,07453	0	0,01176
Искусств_Интеллект__аудит.txt	0,03711	0	0	0	0	0
Методы исследования операций.txt	0,11517	0,04935	0	0,06129	0	0
Моделирование АнализПО.txt	0,00986	0,12825	0,01215	0,01562	0	0,00986
Проектирование ИС_рус.txt	0	0,09512	0,01673	0	0,02151	0,01358
Искусств_Интеллект_менеджмент.txt	0,03793	0	0	0	0	0
СППР.txt	0,03439	0,13757	0	0	0	0
Операционные системы.txt	0	0,17435	0,01431	0	0,03681	0,05811

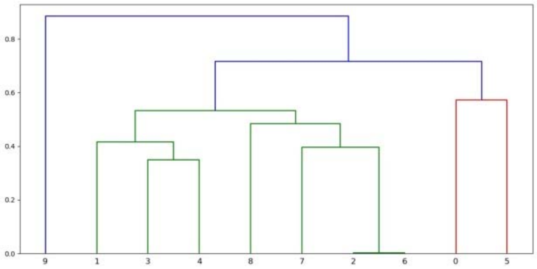
Система управления базами данных Oracle.txt	0	0	0	0	0	0,02851

Now that the task of choosing variables (features) and objects (syllabuses) has been completed, we can proceed to calculating the similarity between syllabuses. As already noted, a measure of cosine similarity of vectors is used. The resulting weight matrix was processed using the cosine_similarity function, which receives a vector weight matrix and returns the cosine distance matrix.

```
        0          1         2          3          4          5          6          7          8          9
0   1.00000   0.36921   0.37148   0.38233   0.32973   0.42872   0.37685   0.32661   0.45771
0.08478
1   0.36921   1.00000   0.45026   0.53988   0.62984   0.13686   0.45433   0.39977   0.36821
0.28624
2   0.37148   0.45026   1.00000   0.47879   0.52054   0.24371   0.99700   0.61441   0.57095
0.09372
3   0.38233   0.53988   0.47879   1.00000   0.65034   0.12133   0.48224   0.60526   0.35365
0.09649
4   0.32973   0.62984   0.52054   0.65034   1.00000   0.14887   0.52359   0.51259   0.45114
0.12563
5   0.42872   0.13686   0.24371   0.12133   0.14887   1.00000   0.25087   0.13255   0.33992
0.15045
6   0.37685   0.45433   0.99700   0.48224   0.52359   0.25087   1.00000   0.59357   0.57572
0.09669
7   0.32661   0.39977   0.61441   0.60526   0.51259   0.13255   0.59357   1.00000   0.40325
0.07888
8   0.45771   0.36821   0.57095   0.35365   0.45114   0.33992   0.57572   0.40325   1.00000
0.02430
9   0.08478   0.28624   0.09372   0.09649   0.12563   0.15045   0.09669   0.07888   0.02430
1.00000
```

Figure 6: Cosine distance matrix

Figure 6 shows the cosine distance matrix obtained during testing from a sample of 10 syllabuses. During the study of the data, it was found that the documents under indices 2 and 6 are the most similar. For a more visual representation, it can be used a graphic image of the calculation results - a dendrogram.



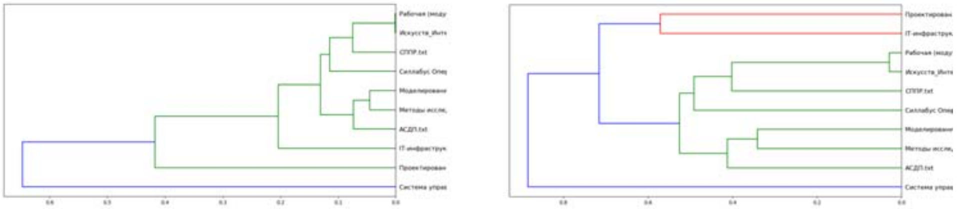


Figure 7: Syllabuses clustering endrogram

For a more accurate representation of the relationship between the disciplines, a graph data model was chosen, which based on a directed graph consisting of nodes and edges. The particularity of the graph model is that it accurately reflects the semantics of the subject area with numerous connections. The next step involves constructing an adjacency matrix using pairwise comparisons of syllabuses.

```
   0  1  2  3  4  5  6  7  8  9
0  0  0  0  0  0  0  0  0  1  0
1  1  0  0  0  1  0  0  0  1  0
2  1  1  0  0  1  1  0  0  1  0
3  1  1  0  0  1  0  0  0  1  0
4  1  0  0  0  0  0  0  0  0  0
5  1  1  0  1  1  0  0  0  1  0
6  1  1  1  0  1  1  0  0  1  0
7  1  1  1  1  1  1  1  0  1  0
8  0  0  0  0  1  0  0  0  0  0
9  1  1  0  1  1  1  0  0  1  0
```

Figure 8: Adjacency matrix

Based on the adjacency matrix, a directed graph was constructed that allows us to represent the relationship between syllabuses. From fig. 9, it can be clearly seen that the vertices correspond to the documents of the analyzed case, and the edges correspond to the relationships between the documents. At the same time, it should be noted that one vertex can have either one or several input edges. It means that all the vertices that belong to one or another connected component belong to the same cluster.

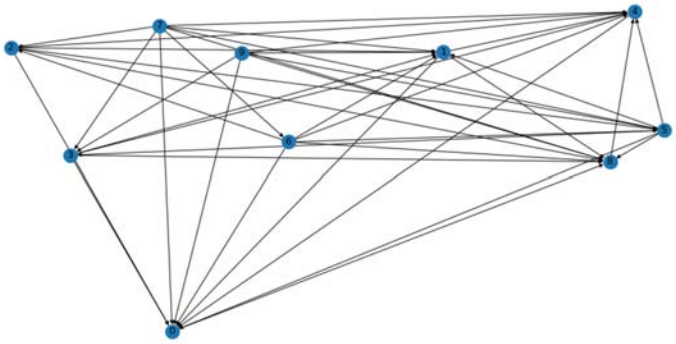


Figure 9: Oriented graph

The next task is the further processing of the oriented graph to build a sequence of studied disciplines. Note that to study a certain discipline, knowledge is necessary as a precondition, i.e. preceding constraints. It can be assumed that the algorithm of topological sorting of the graph will be used, which is represented by topological ordering in compliance with all the restrictions on the preceding.

5. CONCLUSION

An analysis of existing methods for processing text documents showed that there are a number of approaches that are applicable to solve the problem of forming educational programs. In this work, we used the clustering method based on the document's vector model. To obtain vector models and subsequent clustering of documents, a software application was developed in Python using the concept of MVC. It made it possible to identify the corresponding working curricula of disciplines for the formation of professional competencies of educational programs of higher education, as well as to stage-by-stage processing of input data to form a matrix of cosine distances between document vectors. The hierarchical clustering method made it possible to identify syllabuses with the same content, taking into account the context of entities in documents, when automatically extracting entities and correlations between them in the conditions of educational programs of education's subject area without laborious processing and adaptation of knowledge bases. It was necessary to implement such procedures as removing stop words from documents, stemming, determining the importance of a term in the document body by the tf-idf characteristics of the term. The output data can be processed to obtain a visualized representation of graphs, on the basis of which it is possible to identify the sequence of studied disciplines. During the experimental work, 350 educational work programs of disciplines were analyzed for compliance with 120 competencies in the areas of training IT specialists.

An intelligent decision support system created on the basis of the proposed methods and algorithms can be applied in educational institutions to develop new and update existing educational programs considering that the labor market requirements defined by professional standards. This will solve the problem of the professional education content from the demand of employers in the modern conditions of digital economy development.

In the future, the task is to further process the oriented graph to build a sequence of subjects. It is assumed that the algorithm of topological sorting for a graph will be used.

This work was performed as part of project No 598092-EPP-1-2018-1-BG-EPPKA2-CBHE-SP «Modernization of Higher Education in Central Asia through New Technologies».

REFERENCES:

[1] Botov D., Klenin J., "Educational Content Semantic Modelling for Mining of Training Courses according to the Requirements of the Labor Market", *Proceedings of the 1st International Workshop on Technologies of Digital Signal Processing and Storing*, Russia, Ufa, UGATU, 2015. - pp. 214–218.

[2] Yoshitatsu Matsuda, Takayuki Sekiya, Kazunori Yamaguchi, "Curriculum Analysis of Computer Science Departments by Simplified, Supervised LDA", *Journal of Information Processing*, Vol.26, June 2018, pp. 497–508

[3] Botov D.S., Klenin Yu.D., "Approach to Educational Course Comparison Using Natural Language Processing Techniques", *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*, 2017, vol. 17, no. 3, pp. 5–14.

[4] Julius Klenin and Dmitry Botov, "Comparison of Vector Space Representations of Documents for the Task of Matching Contents of Educational Course Programmes", *Ceur Workshop Proceedings*, vol -1975

[5] Bakanova A., Letov N.E., Kaibassova D., Kuzmin K.S., Loginov K.V., Shikov A.N., "The use of Ontologies in the Development of a Mobile E-Learning Application in the Process of Staff Adaptation", *International Journal of Recent Technology and Engineering (IJRTE)*, Volume-8 Issue-2S10, September 2019, pp. 780-789.

[6] Chung H., Kim J., "An Ontological Approach for Ssemantic Modelling of Curriculum and Syllabus in Higher Education", *International Journal of Information and Education Technology*. Vol. 6, no. 5, 2016, pp. 365–369.

[7] Oprea M., "On the Use of Educational Ontologies as Support Tools for Didactical Activities", *Proceedings of the International Conference on Virtual Learning (ICVL2012)*, Nov. 2012, pp. 67–73.

[8] Fedotov A.M., Tussupov J., Sambetbayeva M.A., Sagnayeva S.K., Bapanov A.A., Nurgulzhanova A.N. Yerimbetova A.S., "Using the thesaurus to develop it inquiry systems", *Journal of Theoretical and Applied Information Technology*, Vol. 86, Issue 1, 10 April 2016. pp. 44-61.

[9] Liu Xiaoyong, Croft W Bruce, "Cluster-based retrieval using language models", *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval ACM*, 2004, pp. 186–193.

[10] Sasaki Minoru, Shinnou Hiroyuki, "Spam detection using text clustering", *2005 International Conference on Cyberworlds (CW'05)*, IEEE, 2005, pp. 316-319.

[11] Sergio Decherchi, Simone Tacconi, Judith Redi, "Text clustering for digital forensics analysis". *Computational Intelligence in Security for Information Systems*, Springer, 2009, pp. 29–36.

[12] Dransfield E., Morrot G., Martin J., "The application of a text clustering statisticalanalysis to aid the interpretation of focus group interviews", *Food Quality and Preference* T. 15, № 5, 2004, pp. 477–488.

[13] Orazbayev B., Assanova B., Orazbayeva K., Valentina M. "Problems of multi-criteria optimization in the fuzzy environment and heuristic methods of their solution", *ACM International Conference Proceeding Series, 2019*, pp.78-82.

[14] Feldman R., J. Sanger, "The text mining handbook: advanced approaches in analyzing unstructured data", *Cambridge University Press*, 2007, 410 p.

[15] Moyotl-Hernandez E., Jimenez-Salazar H., "An Analysis on Frequency of Terms for Text Categorization", *Procesamiento del lenguaje natural*, Vol. 33, 2004, pp. 141-146.

[16] Moyotl-Hernandez E., Jimenez-Salazar H., "Some Tests in Text Categorization using Term Selection by DTP", *Proceedings of the Fifth Mexican International Conference on Computer Science ENC'04*., Colima, 2004, pp. 161-167.

[17] Fedotov A.M., Tussupov J., Sambetbayeva M.A., Fedotova O.A., Sagnayeva S.K., Bapanov A.A., Tazhibaeva S.Z., "Classification model and morphological analysis in multilingual scientific and educational information systems", *Journal of Theoretical and Applied Information Technology,* Vol. 86, Issue 1, 10 April 2016, pp. 96-111

[18] Salton G., Buckley C., "Term-weighting approaches in automatic text retrieval", *Information Processing and Management,* 24(5), 1988, pp. 513–523.

[19] Mark S. Aldenderfer, Roger K. Blashfield, *Cluster Analysis*, SAGE Publications, Inc, 1984, 88 p.

[20] "Classification and Clustering", editor by J. Van Ryzin, Academic Press, Inc, New York, San Francisco, London, 1977, pp. 7-19

[21] Boutin F., Hascoet M. Cluster Validity Indices for Graph Partitioning // Proceedings of the Eight International Conference on Information Visualization (IV'04). 2004. – pp. 232-240.

[22] Ying Zhao, George Karypis, "Evaluation of hierarchical clustering algorithms for document datasets", *In Proceedings of the eleventh international conference on Information and knowledge management, ACM, 2002*, pp. 515–524.

[23] Lance G. N., Willams W. T., "A general theory of classification sorting strategies. 1 hierarchical systems", *Computer Journal*, no. 9, 1967, pp. 373–380.

[24] Douglass R Cutting, David R Karger, Jan O Pedersen, John W Tukey. "Scatter/gather: A cluster-based approach to browsing large document collections", *In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval,* ACM, 1992, pp. 318–329.

[25] Michael Steinbach, George Karypis, Vipin Kumar, et al. "A comparison of document clustering techniques", *In KDD workshop on text mining*, volume 400, Boston, 2000, pp. 525–526.

[26] E. Gamma, R. Johnson, R. Helm, J. Vlissides Design Patterns Elements of Reusable Object-Oriented Software – pp. 35-36.

[27] McAuley J. J., Leskovec J., Jurafsky D., "Learning attitudes and attributes from multi-aspect reviews", *In Proceedings of International Conference on Data Mining*, 2012, pp. 1020–1025.

[28] Usama F., Smyth P., Piatetsky–Shapiro G., "From Data Mining to Knowledge Discovery in Databases", *Artifical intelligence Magazine*, 17(3), 1996, pp. 34–54.